An STE-based methodology for constructing Bayesian Belief Networks as applied to the prediction of Economic Indicators

Fernando Torre Mora
Computer Science
University of Missouri
MO, USA

Chanmann Lim
Computer Science
University of Missouri
MO, USA

Adil Al-Azzawi
Electrical and Computer Engineering
University of Missouri
MO, USA

***Abstract*—Instead of treating all random variables as independent set of values and using Discriminative methods for classification tasks. A Bayesian network is a popular graphical representation approach for modeling probabilistic dependencies and causality among a set of random variables to be able to incorporates a huge amount of human expert knowledge about the problem of interest involving diagnostic reasoning. In our study, we set out to construct the Bayesian networks using the standard error for a least-squares linear regression (STE) and the domain knowledge from the literature in the field for predicting the economy.**

Keywords—STE; Bayesian networks; genetic algorithm; domain knowledge; discriminative methods; economic forecasting.

I. INTRODUCTION

Although methods exist to construct Bayesian networks using expert knowledge [1], genetic algorithms [2], and topological ordering [3] few methods exist to construct a Bayesian network using purely mathematical relations between the variables [4]. We believe such a method to be an important contribution to the field, for which reason we set out to develop it.

We based our method on the standard error for a least-squares linear regression, or STE [5]. This metric is consistent with commonly used statisticals such as the correlation coefficient ([6], [7]) and has the additional advantage of allowing us to test causation. This, combined with minimal domain knowledge, allows us to define an unambiguous, valid Bayesian network.

To test our method, we set out to apply it to the problem of predicting economic growth. This problem not only has a large number of variables on which to build on, but it has also become particularly important in the past eight years given the considerable slowdown that has occurred in the global economy. More informed prediction mechanisms would prove invaluable to policy makers and help them make better decisions.

However, It seems unlikely (and in fact is strongly discouraged [8]) that a single model can encompass all the countries in the world accurately. It is necessary, then, to subdivide the countries into regions and build a prediction model for each. This make it an ideal fit to test our Bayesian Network construction methodology.

Our Bayesian Networks aim to show such how significant these factors are to economic growth in each region. We select a series of variables that measure Economy, Production, Education, and Innovation.

We will relate them using STE to create a network, establishing a link where a strong relation is found, and discarding the links that contradict domain knowledge. We will then train and test the networks against the data from the variables.

The next section explains in greater detail the problem of economic growth. Section III describes previous work, both in computing factors of economic growth, and in developing Bayesian Network construction methodologies. Section IV gives the formal problem formulation. Section V explains the complexity of the dataset. Section VI goes into our Bayesian Network construction methodology. Section VII will show how we applied it to the economic prediction problem, followed by a discussion of our results (section VIII), both as far as computed networks and our evaluation of them. We conclude by evaluating our successes and remaining challenges.

II. BACKGROUND THEORY

The global economy is in trouble. Global GDP has been generally falling for nearly a decade (see Fig.1). Governments worldwide and investors have been forced to cut back on spending [10], reducing the strength of the actors that have traditionally been expected to spur economies [9]. A return to the year-to-year growth observed prior to the Great Recession is desirable (the Great Recession can be observed between the years 2007–2009 in Fig.1). This would indicate a return to a global economy capable of withstanding events such as the Asian Financial Crisis (1997–1999 in Fig.1) and the dot-com crash (2001–2003 in Fig.1) without affecting the overall global trend. However, such a strengthening does not seem likely under current conditions. It is only natural for the global question to be how to achieve this.

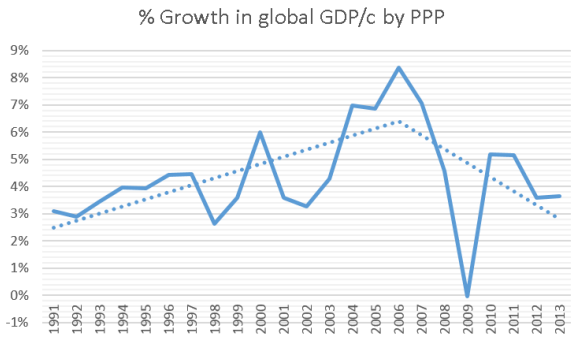

Fig. 1. Percentage of growth in worldwide Gross Domestic Product per capita based on Purchasing Power Parity [17]. The percentage $(y_i - y_{i-1}/y_{i-1})$ is shown with a solid line. The trend (as given by an ordinary linear regression on the years covered) is shown with a dotted line.

Decisions in recent years have seen education funding cut, with widespread opposition. Worldwide, UNESCO has largely led the fight on preventing education funding from being cut. The 2010 report [10] had the explicit aim to develop policy and public awareness on the "social and economic" importance of preparing engineers. The question of what to cut ultimately boils down to how important each of these factors is.

Recent research has shown that investment in education and technology would be extremely favorable to economic growth. As Irina Bokova said in a 2010 UNESCO report,

> The current economic crisis presents challenges and opportunities for engineering. There are encouraging signs that world leaders recognize the importance of continuing to fund engineering, science and technology[. This investment] may provide a path to economic recovery and sustainable development [11].

A country's economic growth has been proven to depend strongly on the number of experts that country has in several areas. ([12], [13]) Specifically, economic growth depends on the knowledge gained by these persons that can be used to manufacture goods, perform services, and improve the productivity of existing processes. This knowledge is known as "productive knowledge" ([14], [10]) and is usually a subset of the knowledge gained by these persons in their higher education studies, or new knowledge generated by them. Therefore, we can measure productive knowledge, and thereby the effect of higher education on the economy, by comparing the number of graduates in different areas, and the research they perform.

However, the exact weight of productive knowledge in comparison to more traditional factors such as government spending [9] has never been quantitatively assessed. Detriments of favoring any one factor exclusively, are well known [15], but the strength of the influence is observed through trial and error, if it is assessed at all.

In this project, we posit that the Bayesian interpretation of conditional probability is a good measure of this influence. To this end, we will construct a Bayesian network to predict two economic indicators from: five production indicators, two education indicators, and three innovation indicators.

III. Previous work

Much has been published linking different economic and education variables to economic growth. In recent years, research has shown that economic development depends strongly on the number of engineers ([10], [13]), scientists ([10], [14], [12]), researchers ([10], [12]), and experts in technology ([12], [13]) While the existence of a strong relation cannot be denied, none of these studies have measured the strength of this link compared to other possible factors mostly because they lack a practical application. A Bayesian Network provides such a practical application

Procedures for constructing Bayesian networks, however, are scant [4]. The most basic method ([1], [3]) consists of the arrangement of variables in cause and effect ordering and the exploitation of conditional independence assumption such that Chain rule can be applied to form the conditional probability table. However, this method is strongly dependent on the ordering of the variables which, in the absence of any true natural order, ends up being pure guesswork

Methodologies have been developed to construct Bayesian networks for specific purposes, mainly using genetic algorithms ([16], [2], [4]). These algorithms are strongly dependent on their initial population which, because they are generated at random, are also pure guesswork.

Comparing variables using purely their statistical properties has been done previously using mutual information as a measure of dependence in [6]; however, this work also points out that this approach does not seek to optimize any statistical, and makes no use of the existing domain knowledge. The authors attempt to introduce an external optimization metric, but notes that it is computationally intensive. In our approach, the variables are compared with a measure of dependence based on an optimized square error.

The authors of [6] also indicate that the number of parents for each node must be restricted in some way, but provide no guidelines on how to do so, leaving the possibility of all other variables to be considered, creating a factorial-order problem. In our approach, we use a domain knowledge graph, thus restricting the number of operations to a polynomial-order problem.

IV. Problem Formulation

To allow future research to perform similar tasks and compare more easily, we present a mathematical formulation presented first as an agent-environment problem and then as a "black box" problem.

An agent charged with the task of predicting development indicators would live in an environment where all the data from all the countries and regions in the world exist. A state in this environment would be defined by the intersection of a year and a region/country. For instance, the state given by ⟨Sub-Saharan Africa, 1999⟩ includes the variables listed in TABLE I.

Such an agent would, taking a selection of these variables, output a prediction for any other variable (For instance, GDP growth). Its goal is to make the correct prediction. We define success of this goal if the prediction is exact, and failure if it is

not. Because we will take the dependences of these variables$(x_1, x_2, \ldots x_n)$, each defined in the discretized domain {High, Medium, and Low}, an exact prediction only needs to be exact if it matches the corresponding discrete variable.

To better understand the problem, we present the black box formulation given in Fig. 2.a: Our economic prediction agents are black boxes that take education, innovation, and production indicators and output economic indicators. However, at a higher level we have the problem of how to generate such a black box. Given that every region is different [8], we should create an agent for each. We therefore define the black box in Fig. 2.b as a black box that outputs agent black boxes from variables and domain knowledge.

V. Dataset

To properly estimate the economy, we need to measure the variables related to productive knowledge ([14], [10]). We used as our data source the World Bank open data bank [17] and hand-picked 11 variables from their list of world development indicators. Each indicator has data from 1960 to 2013, with some values missing. We categorized the indicators into Economic, Innovation, Production, and Education. TABLE II. shows the categories and the indicators' World Bank name.

We choose PPP as a suitable measure of the economy independent of inflation; Industry as a measure of mining, manufacturing, and construction[18]; Unemployment (ILO estimate) since definitions of employment can vary from country to country [19] whereas the ILO uses a single standard definition for all countries[20]; Journal articles as a measure of the amount of research being performed in the region; Trademark applications as a measure of new businesses and products; and Government Expenditure as a measure of how much money the local governments are pumping into their economies, whether it be as incentives or investments.

Labor Force with Secondary Education refers to the number of working-age adults that have completed high school or its local equivalent, while Labor Force with Tertiary Education refers to the number of working-age adults that have completed College or its local equivalent [21].

A. Regional subdivision

Because there are 217 countries and territories in the World Bank, it seems prudent to aggregate them somehow and make use of their combined data. However, this creates the problem of how to perform this aggregation, and how to assign weights to each country. Fortunately, the World Bank also defines 32 aggregations, with the values of each country correctly weighted and added together. We will use the World Bank's seven regions of the world (Fig. 3), which will allow us to cover the world completely [22]. In the cases where a region is divided into "developing only" and "all income levels", we use the latter. Finally, we consider the aggregate for "world" as an eighth region be able to evaluate our accuracy in predicting the global economy.

B. Size

In our dataset we have 4752 data points, 2135 of which are missing, representing 45% missing and 55% not missing.

TABLE I. Sample state of problem environment

Variable name	Percentages
GDP growth	3.6%
Unemployment	8.4%
Services, % GDP	47.9%

TABLE II. Categorization of selected indicators

Categories	Indicators
Economic Indicators	GDP per capita, PPP (constant 2011 international $)
	GDP growth (annual %)
Production indicators	Industry, value added (% of GDP)
	Services, etc., value added (% of GDP)
	Unemployment, total (% of total labor force) (modeled ILO estimate)
Education indicators	Labor force with secondary education (% of total)
	Labor force with tertiary education (% of total)
Innovation indicators	Scientific and technical journal articles
	Trademark applications, total
	General government final consumption expenditure (% of GDP)

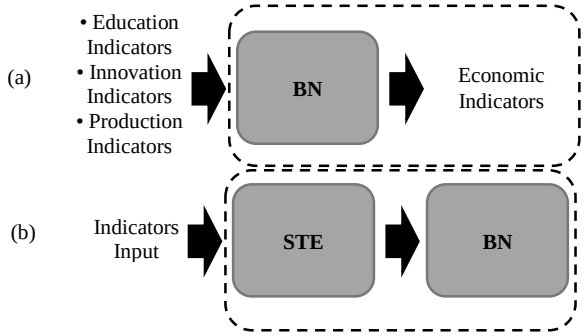

Fig. 2. Black box problem formulation, (a) economic prediction agents, (b) the outputs agent from variables and domain knowledge

Fig. 3. Official world bank regions of the world [23]

The World Bank handles data at country-level granularity. When it performs an aggregate, it leaves the value for that year blank if the data from one-third of the countries in that region are missing [24]. However most regions have very compete data for at least half of the variables. For exact proportions of missing data and dataset dimensions, see TABLE III.

TABLE III. PROPORSION OF MISSING VALUES AND DATASET DIMENSION SIZE

	Regions															
Variables	East Asia		Europe		Latin America		Middle East		North America		South Asia		Africa		World	
Agriculture	***Missing***	11	***Missing***	31	***Missing***	5	***Missing***	22	***Missing***	38	***Missing***	0	***Missing***	5	***Missing***	36
	Not Missing	43	***Not Missing***	23	***Not Missing***	49	***Not Missing***	32	***Not Missing***	16	***Not Missing***	54	***Not Missing***	49	***Not Missing***	18
	%	80	***%***	43	***%***	91	***%***	59	***%***	30	***%***	100	***%***	91	***%***	33
Unemploy-ment	***Missing***	31	***Missing***	31	***Missing***	31	***Missing***	31	***Missing***	31	***Missing***	31	***Missing***	31	***Missing***	31
	Not Missing	23	***Not Missing***	23	***Not Missing***	23	***Not Missing***	23	***Not Missing***	23	***Not Missing***	23	***Not Missing***	23	***Not Missing***	23
	%	42	***%***	42	***%***	42	***%***	42	***%***	42	***%***	42	***%***	42	***%***	42
Tertiary education	***Missing***	54	***Missing***	40	***Missing***	48	***Missing***	54	***Missing***	49	***Missing***	51	***Missing***	54	***Missing***	54
	Not Missing	0	***Not Missing***	14	***Not Missing***	6	***Not Missing***	0	***Not Missing***	5	***Not Missing***	4	***Not Missing***	0	***Not Missing***	0
	%	0	***%***	26	***%***	11	***%***	0	***%***	9	***%***	7	***%***	0	***%***	0
Secondary education	***Missing***	54	***Missing***	40	***Missing***	48	***Missing***	54	***Missing***	49	***Missing***	51	***Missing***	54	***Missing***	54
	Not Missing	0	***Not Missing***	14	***Not Missing***	6	***Not Missing***	0	***Not Missing***	5	***Not Missing***	4	***Not Missing***	0	***Not Missing***	0
	%	0	***%***	26	***%***	11	***%***	0	***%***	91	***%***	7	***%***	0	***%***	0
GDP growth	***Missing***	1	***Missing***	1	***Missing***	1	***Missing***	9	***Missing***	1	***Missing***	1	***Missing***	1	***Missing***	1
	Not Missing	53	***Not Missing***	53	***Not Missing***	53	***Not Missing***	45	***Not Missing***	53	***Not Missing***	53	***Not Missing***	53	***Not Missing***	53
	%	98	***%***	98	***%***	98	***%***	83	***%***	98	***%***	98	***%***	98	***%***	98
GDP per capita, PPP	***Missing***	30	***Missing***	30	***Missing***	30	***Missing***	30	***Missing***	30	***Missing***	30	***Missing***	30	***Missing***	30
	Not Missing	24	***Not Missing***	24	***Not Missing***	24	***Not Missing***	24	***Not Missing***	24	***Not Missing***	24	***Not Missing***	24	***Not Missing***	24
	%	44	***%***	44	***%***	44	***%***	44	***%***	44	***%***	44	***%***	44	***%***	44
Gov. final consumption	***Missing***	0	***Missing***	0	***Missing***	0	***Missing***	8	***Missing***	0	***Missing***	0	***Missing***	0	***Missing***	0
	Not Missing	54	***Not Missing***	54	***Not Missing***	54	***Not Missing***	46	***Not Missing***	54	***Not Missing***	54	***Not Missing***	54	***Not Missing***	54
	%	100	***%***	100	***%***	100	***%***	85	***%***	100	***%***	100	***%***	100	***%***	100
Services	***Missing***	11	***Missing***	31	***Missing***	5	***Missing***	22	***Missing***	38	***Missing***	0	***Missing***	5	***Missing***	36
	Not Missing	43	***Not Missing***	23	***Not Missing***	49	***Not Missing***	32	***Not Missing***	16	***Not Missing***	54	***Not Missing***	49	***Not Missing***	18
	%	80	***%***	43	***%***	91	***%***	59	***%***	30	***%***	100	***%***	91	***%***	33
Industry	***Missing***	11	***Missing***	31	***Missing***	5	***Missing***	22	***Missing***	38	***Missing***	0	***Missing***	5	***Missing***	36
	Not Missing	43	***Not Missing***	23	***Not Missing***	49	***Not Missing***	32	***Not Missing***	16	***Not Missing***	54	***Not Missing***	49	***Not Missing***	18
	%	80	***%***	43	***%***	91	***%***	59	***%***	30	***%***	100	***%***	91	***%***	33
Scien. & tech. journal articles	***Missing***	26	***Missing***	30	***Missing***	28	***Missing***	28	***Missing***	28	***Missing***	26	***Missing***	28	***Missing***	26
	Not Missing	28	***Not Missing***	24	***Not Missing***	26	***Not Missing***	26	***Not Missing***	26	***Not Missing***	28	***Not Missing***	26	***Not Missing***	28
	%	52		44		48		48		48		52		48		52
Trademark application	***Missing***	11	***Missing***	32	***Missing***	8	***Missing***	15	***Missing***	0	***Missing***	4	***Missing***	54	***Missing***	20
	Not Missing	43	***Not Missing***	22	***Not Missing***	46	***Not Missing***	39	***Not Missing***	54	***Not Missing***	50	***Not Missing***	0	***Not Missing***	34
	%	80	***%***	41	***%***	85	***%***	72	***%***	100	***%***	93	***%***	0	***%***	63

VI. METHOD

Our main contribution to the field lies in our methodology, which is summarized in Fig. 4. This methodology is fully automatable and can be adapted to any domain.

We start off by selecting and categorizing the variables. A simple linking of the categories using the domain knowledge creates a graph which we term our Domain Knowledge Model, which allows the procedure to be readily be applied to other domains by simply changing the variables involved. We then calculate the degree of dependence between all the variables in every pair of linked categories, after some minimal preprocessing. By using just the links between categories, as opposed to comparing all against all, This reduces our computations from ***(n+m)!*** (similar to the approaches used by [6], [1], and [3]) to ***n×m***; where *n* and *m* are the number of variables in each category. The resulting Bayesian network can then be trained and evaluated normally.

A. Preprocessing

1) Scaling

Most variables in our dataset are percentages. However, the variables for journal articles are numerical quantities. It is good practice to train Bayesian networks with normalized values, all within the same range, for which reason we perform a simple scaling.

Simply we take the value for each of these variable according to its region and divide it by the population of this region times 100. This is formally stated in (1) where R refers to each region in the dataset

$$\text{scaled\_variable}_{\text{R}} = \frac{\text{old\_variable}_{\text{R}}}{\text{Population}_{\text{R}}} \times 100 \quad (1)$$

2) Missing value treatment

The missing values were simply ignored. Since each variable has a deep complex economic implication defined solely by The World Bank[24], and they highly depend on many other evidences, we decided that the prediction of those missing values by filling in the best values or with distribution using EM algorithm would be a crude estimation if not biased towards the low amount of data in our study.

We select which rows to ignore in each operation using matlab's *isnan* function.

B. Bayesian Network Construction

In our work, we used the standard error for a least-squares linear regression or STE [5], [13] to calculate dependence. We note that the methodology is not tied to this statistical (Friedman et al. [6] suggests correlation or mutual information for this purpose; however, STE is known to be consistent with both of these statisticals). We use STE because it gives an indication of which variable is the dependent variable and which one is the independent variable. Specifically, a small STE(Y, X) implies a strong causative relation where Y depends on X [6]. The formula for STE is shown in Equation (2) with Y and X being vectors of values that have a length n, and with $\bar{Y}$ and $\bar{X}$ being their respective sample means.

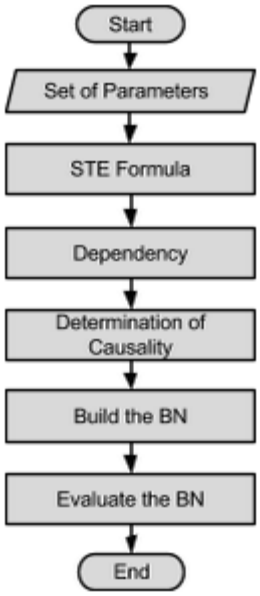

Fig. 4. Proposed System Flowchart

$$\text{STE}(Y,X) = \sqrt{\frac{1}{n-2}\left(\sum_{y\in Y}(y-\bar{Y})^2 \; - \frac{\left(\sum_{x\in X, y\in Y}(x-\bar{X})(y-\bar{Y})\right)^2}{\sum_{x\in X}(x-\bar{X})^2}\right)} \quad (2)$$

To make the result easier to interpret, we use Equation (3) from [13] so that higher values are better. This equation also normalizes the measure into the [0,1] range for better readability. We call this the degree of **dependency**. Note that vertical bars denote absolute value.

$$\text{dependency}(Y,X) = 1 - |\text{STE}(Y,X)|/\bar{Y} \quad (3)$$

If $\text{dependency}(Y,X) > \text{dependency}(X,Y)$, we conclude there may be a causal relationship between X and Y, with X being the cause and y being the effect, and thus add an arc from X to Y in our Bayesian network. However, if $\text{dependency}(Y,X) < \text{dependency}(X,Y)$, rather than concluding Y is the cause and X is the effect, we discard it entirely. This is because the comparison is made following the links in the Domain Knowledge Model, and adding such an arc would contradict the domain knowledge.

Because data may be prone to errors, outliers, or may simply not be complete enough, we define the case when $\text{dependency}(Y,X)$ is *slightly less* than $\text{dependency}(X,Y)$. This is the case when $\text{dependency}(Y,X) \leq \text{dependency}(X,Y)$, but are close enough to consider that, given slightly better data, we could have $\text{dependency}(Y,X) > \text{dependency}(X,Y)$. We define a **threshold** of three percent as the limit of this closeness; however, this threshold is user-defined as any number between zero and one and only depends on the desired number of arcs. Note that we can similarly use simple STE (2) instead of dependency, but in this case the threshold would have to be defined between $-\bar{Y}$ and $\bar{Y}$. In other words, the normalization of STE in (3) means that, when the threshold values are interpreted as maximum error, they are expressed in means of Y.

```
Function GET_STE_VALUES (child_layer, parent_layer)
    Inputs:
        child_layer   ← set of vectors, values of each of the
                        believed dependent variables
        parent_layer  ← set of vectors, values of each of the
                        variables the members of child_layer are
                        believed to depend on
    Outputs:
        return DEPENDENCY_TABLE(
            DEPENDENCY_VALUES(child_layer, parent_layer)
            .03, .6 )
```

```
Function DEPENDENCY_VALUES (child_layer, parent_layer)
    Outputs:
        for i ∈ parent_layer
            for j ∈ child_layer
                V[i to j]  ← DEPENDENCY(i,j)
                V[j to i]  ← DEPENDENCY(j,i)
            end
        end
    return V
```

```
Function DEPENDENCY_TABLE (V, threshold, minimum)
    Outputs:
        for (i to j) ∈ V
            if V[i to j] ≥ minimum
                if V[i to j] > V[j to i])
                    Graph.add_arc(i to j)
                else
                    if V[j to i] − V[i to j] < threshold
                        Graph.add_arc(i to j)
                    end
                end
            end
        end
    return Graph
```

Fig. 5. Pseudocode for the Bayesian Network Construction algorithm

```
Function LEARNING(Dataset, bn) returns cpt of all variables in
the Bayesian network
    Input: Dataset, the dataset (already discretized)
        bn, the Bayesian network
    Outputs:
        X ← bn.Vars /* All variables in the Bayes net */
        Q(X) ← a distribution over X, initially empty
        for each x_i of X do
            if parent(x_i) is empty then
                    Q(X) ← PR(x_i |Dataset)
            else
                    Q(X) ← CPT(x_i, parent(x_i)|Dataset)
            end
        end
        return Q(X)
```

```
Function PR(X, d) returns probability of X given the domains is d
    Input: X, the data of a random variable
        d, the domains of X
    Outputs:
        k ← GET_LAPLACE_K();
        Q(X) ← a distribution over X, initially empty
        for each value of d do
            Q(X) ← (COUNT(X == value) + k) / (COUNT(X) + k *
                    LENGTH(d))
        end
        return Q(X)
```

```
Function CPT(X, e) returns probability of X given the evidence e
    Input: X, the data of a random variable
        e, the evidence variables
    Outputs:
        domain ← GET_DOMAINS();
        Q(X) ← a distribution over X, initially empty
        for each value of domain do
            Q(X) ← PR(X, value)
        end
        return Q(X)
```

Fig. 6. Pseudocode for parameter learinng algorithm

Finally, we define the case when dependency(Y, X) is simply too small to imply a causal relationship. Since we do not want to add an arc in these cases even if dependency$(Y, X) >$ dependency(X, Y), we discard the results entirely. We define a **minimum** of 60% as the measure of this smallness. Again, this minimum is user-defined between zero and one and only depends on the desired number of arcs, and again, one can use simple STE, but the range becomes a function of $\overline{Y}$.

The result is a Bayesian network graph. Any graph representation can be used. In our work, we used a simplified adjacency matrix *T* where only nodes that could have children, as given by the domain knowledge, were columns and only nodes that could have parents, as given by the domain knowledge, were rows. That is, if the Domain Knowledge Model is seen as a graph *K*, we omit the source nodes of *K* from the rows of *T* and the sink nodes of *K* from the columns of *T*.

The network construction algorithm is summarized in Fig. 5.

1) Complexity analysis

The runtime of the construction algorithm depends strongly on the Domain Knowledge Model and how many variables it receives.

In the best case, each category will have exactly one variable, which would imply the Bayesian network structure is already known, and merely needs to be simplified. In this case, the algorithm performs $2m$ operations (dependency(Y, X) and dependency(X, Y)), where m is the number of arcs in the Domain Knowledge Model. Since a Bayesian network must be a directed acyclic graph, this case may have m being anywhere between $n - 1$ (Markov chain) and $n(n-1)/2$ (transitive closure of a fully reachable graph), where n is the number of variables. Therefore, the algorithm is $\Omega(n)$ and $O(n^2)$ in the best case. This is comparable to the best case in [6] where each node has one or two candidate parents.

In the worst case, each category has the same number of variables: n/c where c is the number of categories and n is a multiple of c such that $n \geq 2c$. To evaluate each arc, the members of each category in the arc's source have to be compared with the member of each category in the arc's sink, each comparison of which requires two operations, or $2(n/c)^2$

per arc between categories for a total of $2m(n/c)^2$. Since, again, there may be anywhere between $n-1$ and $n(n-1)/2$ arcs between categories, the algorithm is $\Omega(n^2)$ and $O(n^4)$ in the worst case. This is much better than the worst case in [6] where all other nodes are candidate parents, leading to $O(n!)$.

It should be noted that, given the nature of Bayesian networks as inference engines, the worst case is highly unlikely to be encountered in practice. It is more likely that there will be a category with much less variables than the others, since there is always a small group of target variables (usually one). For this reason we can assume an average order of n^2.

C. Bayesian Network Evaluation

We manually build a Bayesian network designed to handle discrete values and thus learn the Conditional Probability Tables for the network. Inference is then performed using elimination or enumeration on the learned probabilities.

For comparison purposes, we define a Baseline Structure, consisting of the joint probability of all variables – in effect, a Domain Knowledge Model with just two categories: One containing the target variable(s), and one containing all others.

1) Discretization

To perform the conversion, values are discretized into High, Medium, or Low using Equation (4), where x is the specific value being converted; X is the multiset of all the values the variable takes on in the dataset; H, M, L represent High, Medium, or Low respectively; $m(x)$ is the Maximum Likelihood estimator for the mean, given by $(\sum_{x\in X} x)/|X|$; and $d(X)$; is the Maximum Likelihood Estimator for the standard deviation, given by Equation (5). Note that here, the vertical bars denote the cardinality of the set.

$$\text{discretize}(x \in X) = \begin{cases} H \text{ if } x > m(X) + d(X) \\ L \text{ if } x < m(X) - d(X) \\ M \quad \text{otherwise} \end{cases} \tag{4}$$

$$\sqrt{\frac{1}{|X|}\sum_{x\in X}\left(x - m(X)\right)^2} \tag{5}$$

2) Baseline

In the Baseline structure, the joint probability of everything is computed (all the variables). In this model, all variables directly affect the target (the economic indicators in our work).

3) Learning

In the learning part, the Bayesian network parameters have been learned by computing the conditional probabilities through Maximum Likelihood. By using the formula in Equation (6),

$$\begin{aligned} &P(x_1, \dots, x_n) \\ &= \prod_{i=1}^{n} P\left(x_i \middle| \text{parent}(X_i)\right) \\ &= P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1)P(x_5|x_2, x_3, x_4) \end{aligned} \tag{6}$$

4) Inference

We have experimented with the prediction of the economic indicators using two exact inference algorithms, which may be used interchangeably.

```
Function PREDICT_BY_ELIMINATION(e, factors) returns
prediction of PPP variable
    Input: e, the evidences
           factors, the joint factors
    Outputs:
        probability ← factors
        for each variable ∈ e do
            if variable is missing then
                probability ←   add all values in probability for
                                this variable
            else
                probability ←   Lookup all values in probability
                                where this variable = variable ∈ e
            end
        end
        prediction ← MAX_INDEX( probability)
    return prediction
```

Fig. 7. Pseudocode for inference algorithm (Elimination)

```
Function PREDICT_BY_ENUMERATION(t, P, e) returns
prediction of PPP variable
    Input: t, the target variable
           e, the known evidences
           P, conditional probabilities
    Outputs:
        for each parent of t
            if parent ∈ e not missing
                joint[parent] ← P(parent = parent ∈ e)
            else
                joint[parent] ← PREDICT_BY_ENUMERATION(
                                parent, P, parents(parent) ∈ e )
            end
        end
        for each possible value ∈ t
            probability[possible value] ← ∑P(possible value|e) ×
                                          ∏joint
        end
    return MAX(probability)
```

Fig. 8. Pseudocode for inference algorithm (Enumaration)

In Fig. 7 we give the elimination inference algorithm which accepts a set of evidences and the joint factors of all variables then checks if an evidence variable is hidden to sum out all of its possible values otherwise just lookup the probability from the probability distribution. In Fig. 8, we give the enumeration algorithm. These algorithms are run iteratively over all the data samples.

5) Accuracy

To define our accuracy, which is success if the prediction is exact, and failure if it is not, we will take the dependences of these variables $(x_1, x_2, \dots x_n)$, each defined in the discretized domain {High, Medium, and Low}. We recall that an exact prediction only needs to be exact if it matches the corresponding discrete variable or not.

Mathematically, we define it as the number of predicted values that match the actual values divided by the total number of known values. This is summarized in equation (7). Note that here vertical bars denote cardinality.

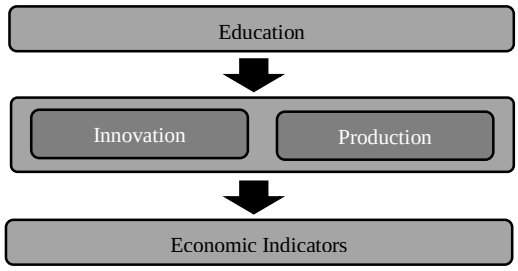

Fig. 9. Three-layer Domain Knowledge Model for the Development Indicators problem

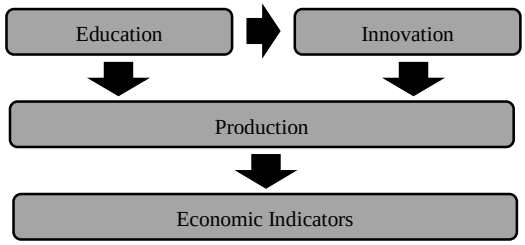

Fig. 10. Alternative Domain Knowledge Model for the Development Indicators problem

$$\text{accuracy} = \frac{|\{i : x_i \in \text{predicted} \wedge y_i \in \text{actual} : x_i = y_i\}|}{|\{y : \text{actual} : y \text{ is not missing}\}|} \quad (7)$$

VII. Proposed Solution

A. Development Indicators Domain Knowledge Model

To build the Domain Knowledge Model, consider the categories from TABLE II. Previous work has shown that relationships between these broad categories is known: Education affects Innovation and Production ([10], [12], [13]); Innovation and Production affect the Economy ([9], [14]). Education is known to not have a direct effect on the economy due to the necessity of applying productive knowledge to Innovation and Production for its effects to become visible ([10], [14]). This gives us a three-layer structure (Fig. 9).

Modelling the domain knowledge also allows us to delimit the number of dependency values we would have to calculate. Suppose we have six variables. If we were to compare all variables against all others (worst case in [6]), we would need to perform 6! comparisons, or calculate 1440 dependency values. Following the three layer structure, with four variables in the middle layer and two in each of the others, we only need to compute 16 dependency values. We note that, although this is in the order of 6^2, it is much less than 6^2.

As evidence of the versatility of our method, we consider a different Domain Knowledge Model. In this model, Innovation also does not affect the economy directly, but rather must also be applied to production first. However, we still allow Education to affect Innovation. The resulting model can be seen in Fig. 10.

Again, this model delimits the number of dependency values we would have to calculate. Using all 11 variables selected in TABLE II. , we would have to perform 11! comparisons, or compute 79.8 million dependency values. With the given structure, the four variables in the middle layer are compared against seven other variables, while the education variables are compared against three others, yielding a grand total of 62 dependency values: much less than 11! or indeed $11^{2.}$.

VIII. Results

A. Bayesian Network structures

1) Bayesian Network 1 (BN1)

We constructed nine networks for our first belief system: a baseline as described in section VI.C.2), and a network with the dependency analysis results for each of the eight regions. The resulting networks are shown in Fig. 11. We only considered six of the selected variables to construct these networks: Tertiary education, Agriculture, Industry, Government spending, Journal articles, and GPD by PPP. We keep our previously established categories for these variables.

In most regions where Tertiary education is considered, it is found to be linked to industry and innovation, but not to agriculture. This makes intuitive sense. Latin America is the exception, but it is known to traditionally have placed higher emphasis on using its tertiary institutions to improve agriculture than to perform research [27]. The service sector was left out of all networks which again make sense because service-based economies are a very recent development [25] and our data spans 54 years.

2) Bayesian Network 2 (BN2)

We constructed eight networks for our second belief network: one for the dependency analysis results for each of the eight regions. The resulting networks are shown in Fig. 12. We considered all 11 of the selected variables to construct these networks using the previously established categories.

When adding secondary education, there generally isn't a connection to the innovation variables except in Europe. Once again, this makes sense given students first encounter with journal articles generally occurs at the university level [1] and Europe, due to its historical diversity of educational systems, has great overlap between tertiary and secondary education.

The networks also make sense when compared to each other: BN1 establishes, for example, that the global economy depends strongly on just agriculture and journal articles (Fig. 11.i). However, because journal articles do not affect agriculture strongly, it disappears from its corresponding network in BN2 (Fig. 12.g). Similarly, the economy for South Asia (Fig. 11.g) is found to depend strongly on journal articles. When we move it to a higher level, we find no factors that affect the economy, and therefore have no network for South Asia in BN2.

We are similarly unable to find any variables that affected GDP growth, for which reason it is absent from all networks. We believe this is because GDP growth is the only variable that measures change from year to year.

B. Evaluation results

We present our results first for the Baseline Belief Network, which was run for the data from each of the regions. We next present our results for the network specially computed for each of the regions, run on that regions' data.

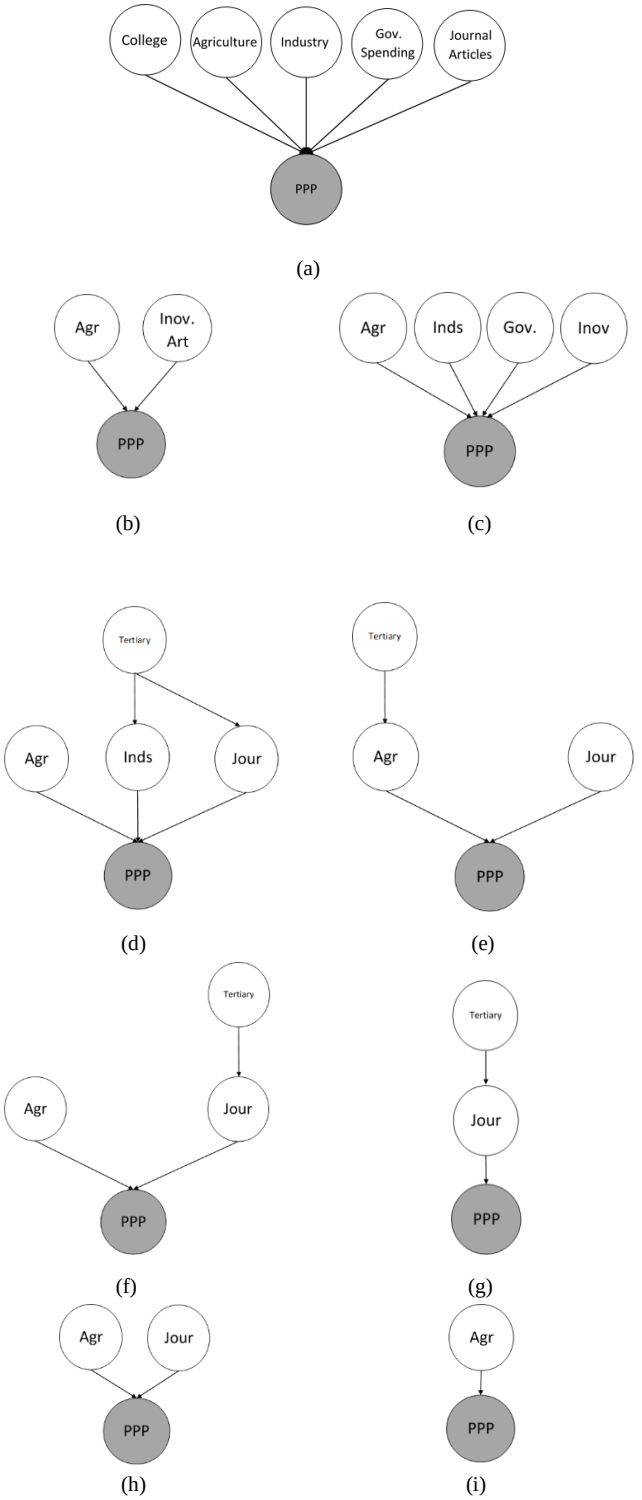

Fig. 11. Beleif Network 1 structures for (a) the baseline (used for comparison purposes), (b) Sub-Saharan Africa region, (c) Middle East and North Africa region, (d) Europe and Central Asia region, (e) Latin America region, (f) North America region, (g) South Asia region, (h) East Aisa and Pacific region, (i) World; where "Agr" represents Agriculture, "Innov. Art" and "Jour" represent Scientific and Journal Articles, "Inds" represents Industry, "Gov" represents Government final consumption expenditure, "Tertiary" represents Labor force with tertiary education, and "PPP" represents GDP per capita by Purchasing Power Parity

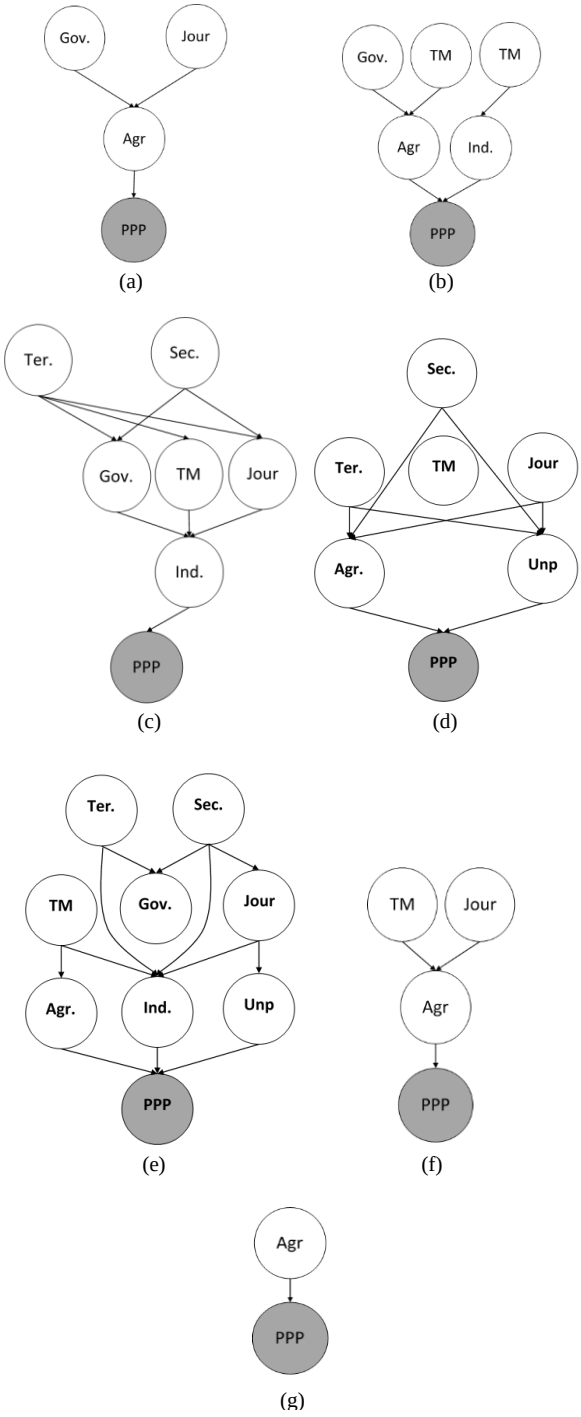

Fig. 12. Beleif Network 2 structures for (a) Sub-Saharan Africa region, (b) Middle East and North Africa region, (c) Europe and Central Asia region, (d) Latin America region, (e) North America region, (f) East Asia and Pacific region, (g) World; where "TM" repesents Trademark Applications, "Ter" reprsents Labor force with tertiary education, "Sec" represents Labor force with secondary education, an "Unp" represents unemployment.

We run the Baseline Belief network and the network for "Middle East and North Africa" using the Elimination algorithm. We run all other networks using the Enumeration algorithm.

TABLE IV. ACCURACY RESULT FOR THE BASELINE BELIF NETWORK STRUCTURE

Region	Accuracy
Africa	0.25
Middle East	0.25
Europe	0.92
Latin America	0.58
North America	0.46
East Asia	0.20
South Asia	0.58
World	0.20

TABLE V. ACCURACY RESULT FOR THE BASELINE BELIFE NETWORK STRUCTURE

Region	Accuracy
Africa	0.67
Middle East	0.75
Europe	0.79
Latin America	0.71
North America	0.54
East Asia	0.67
South Asia	0.65
World	0.55

1) Baseline Structure

As described above, the baseline structure is the joint probability for all variables affecting PPP. The highest inference accuracy is for "Europe and Central Asia" with 92% accuracy and the lowest one is the World with 20% accuracy. The accuracy of the baseline structure for each region is shown in I

2) Belief Network No.1 Structure for World Regions

As with the Baseline structure, the highest inference accuracy is for the "Europe and Central Asia region"; however, the accuracy for the computed network is of 79% accuracy which is lower than the baseline by 13%. The lowest inference value for BN1 is in the North America with 54%. This is better than the baseline, where the accuracy was 46%. We also improve on the accuracy of the world, which was the lowest for the baseline. We show the accuracy for each computed region in IA comparison of the baseline and the networks computed for each region is shown in Fig. 13.

I. CONCLUSIONS

We were able to create logical models for all selected regions using our methodology with two different Domain Knowledge Models. The resulting models are consistent with the knowledge known about the regions during the years covered by the data.

Our networks in general provide accuracy improvements over the baseline. Our baseline provided us an accuracy of 20% to 58% in seven out of eight regions, including the aggregate for "World", while the Bayesian networks generated by our first Domain Knowledge Model improved that accuracy to 54% to 75% in the same regions.

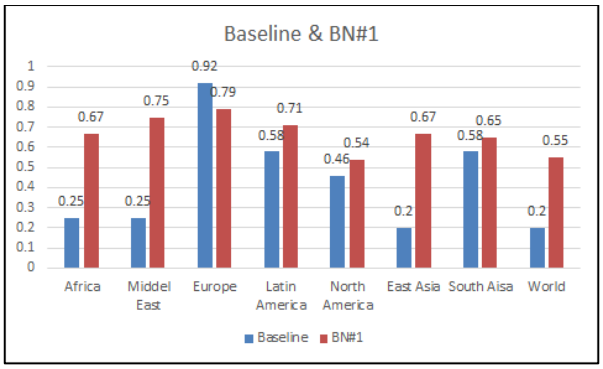

Fig. 13. Performace comparsion between Baseline Belife Networks and the computed Beleif Network 1

For Europe, we were not able to improve on the accuracy of the baseline (92 percent). We suspect this may be due to insufficient data (the aggregation caused the existence of too many missing values) or because Europe is inherently exceptional.

We were similarly unable to construct networks to determine GDP growth for all regions, or GDP per Capita by PPP in South Asia. Better data, as well as variables that measure year-to-year changes, are needed to fully determine whether this methodology is adequate for these cases.

A. Future Work

In future analysis, we would like to reduce our proportion of missing values to better evaluate our Bayesian networks. One of the main reasons our proportion of missing values was so high was because of how the World Bank aggregates regions and the way its regions are defined. One way to reduce this proportion is to aggregate regions differently, or change how missing data is handled during aggregation [24].

Bayesian networks are also capable of handling multiple queries other than just the target variable. We would like to evaluate the accuracy of questions like:

- What does a strong economy and a weak education system imply for the production sectors?
- How high must education be in each region for a high GDP?
- What is the probability the GDP will drastically change given how we know current events affect other indicators?

We also have, so far, manually implemented each of the Bayesian networks. We are aware that this process is automatable, especially given that our methodology generates a graph adjacency matrix. We would like to experiment with different Domain Knowledge models and see their effect on the accuracy. Similarly, we would like to use more variables from the World Bank to see their effect on the accuracy.

In addition, we do not yet have a mathematically proven estimate on the effects of tuning the "minimum" and "threshold" parameters to have on the accuracy. We would like to perform more experiments varying these and see the resulting Bayesian networks.

Finally, we are aware that the data is temporal in nature, and our models have not incorporated this into their consideration. While the methodology is flexible enough to add categories related to the previous time slice, we have not yet experimented with its effects on the accuracy. This, likes other items in this section, are left as future work.

REFERENCES

[1]. Russell, S, and Norvig, P. (2010) "A method for constructing Bayesian Networks." *Artificial Intelligence: A Modern Approach. Third Edition*. §14.2.1.1. Pearson Education.

[2]. Shapcott, M., Sterritt, R., Adamson, K., and Curran, E. (1999) "NETEXTRACT - Extracting Belief Networks in Telecommunications Data." The Pennsylvania State University. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.6865.

[3]. Chickering, D. and D. Heckerman. (1997) "Scalable Methods for Learning Bayesian Networks". Microsoft Corporation (assignee). Patent 7,251,636. http://patentimages.storage.googleapis.com/pdfs/US7251636.pdf

[4]. Cooper, G. and Herskovits, E. (1992) "A Bayesian Method for the Induction of Probabilistic Networks from Data." *Machine Learning*. Volume 9, Issue 4, pp 309-347. Springer International. http://link.springer.com/article/10.1007%2FBF00994110

[5]. Microsoft. (2007). Excel statistical functions: STEYX. *Microsoft Knowledge Base*. https://support.office.com/article/STEYX-function-4cb00b43-c209-4509-980b-ce4ec8431897.

[6]. Friedman, N., Nachman, I., Peér, D., (1999) "Learning of Bayesian Network Structure from Massive Datasets: The 'Sparse Candidate' Algorithm." *UAI'99 Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp 206-215. http://dl.acm.org/citation.cfm?id=2073820.

[7]. Zady, M. "Correlation and Simple Least Squares Regression" (Aug 2000) Westgard QC. https://www.westgard.com/lesson42.htm.

[8]. Migiro, A. (Jun 16 2010) "No 'one size fits all' approach to development." UN News Center. http://www.un.org/apps/news/story.asp?NewsID=35048.

[9]. Drahokoupil, J. "Investment incentive." (May 30 2013) *Encyclopædia Britannica Online*. Encyclopædia Britannica Inc, http://www.britannica.com/EBchecked/topic/1929166/investment-incentive.

[10]. UNESCO. (2010). *Engineering: Issues, Challenges and Opportunities for Development UNESCO Report.* Unesco. http://unesdoc.unesco.org/images/0018/001897/189753e.pdf.

[11]. Bokova, I in UNESCO. (2010). "Foreword." *Engineering: Issues, Challenges and Opportunities for Development* UNESCO Report. http://unesdoc.unesco.org/images/0018/001897/189753e.pdf

[12]. Jaffe, K., Rios, A., and Florez, A. (2012). Statistics shows that economic prosperity needs both high scientific productivity and complex technological knowledge, but in different ways. *Interciencia* vol.38. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2171464

[13]. Mora, J., Torre, F., and Torre, F. (2013) "Contribución de la enseñanza de la ingeniería a la generación de conocimiento productivo". *Memorias del IV Congreso Iberoamericano de Enseñanza de la Ingeniería. Asociación Iberoamericana de Instituciones Enseñanza de la Ingeniería*. Barquisimeto, Lara, Venezuela. ISBN: 978-980-6526-01-3.

[14]. Hausmann, R., Hidalgo, C., Bustos, S., Coscia, M., Chung, S., Jiménez, J. S.,Simoes, A., and Yıldırım, M.. (2011). *The Atlas of Economic Complexity - Mapping paths of prosperity*. Boston: Center for International Development - Harvard University.

[15]. ITEP. (2013) "Tax Incentives: Costly for States, Drag on the Nation." *ITEP Reports*. Institute on Taxation and Economic Policy. http://itep.org/itep_reports/2013/08/tax-incentives-costly-for-states-drag-on-the-nation.php.

[16]. Helman, P., Veroff, R., Atlas, S., and Willman, C. (January 20, 2005) "A Bayesian Network Classification Methodology for Gene Expression Data." *Journal of Computational Biology*, Vol. 11, Issue 4. http://online.liebertpub.com/doi/abs/10.1089/cmb.2004.11.581 .

[17]. World Bank. (2015) *World Bank Open Data Bank*. World Bank Group. http://databank.worldbank.org.

[18]. "Industry, value added (% of GDP)." The World Bank. http://data.worldbank.org/indicator/NV.IND.TOTL.ZS.

[19]. "Unemployment, total (% of total labor force) (national estimate)." The World Bank. http://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS

[20]. International Labour Organization. "Main statistics (annual) - Unemployment" *LABORSTA Internet* http://laborsta.ilo.org/applv8/data/c3e.html

[21]. International Labour Organization. "Indicator 8: Educational Attainment of the Youth Labour Force." *Youth Labour Market Indicators*. Youth Employment Network. http://www.ilo.org/public/english/employment/yen/whatwedo/projects/indicators/8.htm

[22]. World Health Organization. "Definition of region groupings" http://www.who.int/healthinfo/global_burden_disease/definition_regions/en/

[23]. Han-teng Liao. (2014) "World Bank region Natural Earth". *Wikimedia Commons*. Oxford, United Kingdom. http://commons.wikimedia.org/wiki/File:World_Bank_region_Natural_Earth_en.png

[24]. World Bank. (2015) "Methodologes." *World Bank Open Data Bank*. World Bank Group. http://data.worldbank.org/about/data-overview/methodologies

[25]. Piñeiro, M, and Trigo, E. (1977) *La transferencia de ciencia y tecnología y la educación agrícola*. Instituto iberoamericano de Ciencias Agrícolas. Organization of American States. Bogotá, Colombia.

[26]. Cali, M., Ellis K., and te Velde (2008) "The contribution of services to development: The role of regulation and trade liberalisation" London: Overseas Development Institute

[27]. Anderson, C. (2012) "Can high school students read primary research papers?" Genegeek. http://genegeek.ca/2012/12/reading-scientific-papers/